Visemic processing in audiovisual discrimination of natural speech

Dubois, C ; Otzenberger, H ; Gounot, D ; Sock, R ; Metz-Lutz, M-N

DOI: https://doi.org/10.1016/j.neuropsychologia.2012.02.016

Other titles: Visemic processing in audiovisual discrimination of natural speech: A simultaneous fMRI–EEG study

Neuropsychologia 50 (2012) 1316–1326

Contents lists available at SciVerse ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Visemic processing in audiovisual discrimination of natural speech: A simultaneous fMRI–EEG study

Cyril Dubois[a,*], Hélène Otzenberger[b], Daniel Gounot[b], Rudolph Sock[a], Marie-Noëlle Metz-Lutz[b]

[a] Institut de Phonétique de Strasbourg, Équipe de Recherche Parole et Cognition, U.R. 1339 – LILPA, Université de Strasbourg, France
[b] Laboratoire d'Imagerie et de Neurosciences Cognitives, LINC, UMR 7237 CNRS/Université de Strasbourg, France

ARTICLE INFO

Article history:
Received 3 March 2011
Received in revised form 28 January 2012
Accepted 22 February 2012
Available online 1 March 2012

Keywords:
Lipreading
Visual distinctiveness
Syllables
Phonetic features
Motor cortex
MT+V5

ABSTRACT

In a noisy environment, visual perception of articulatory movements improves natural speech intelligibility. Parallel to phonemic processing based on auditory signal, visemic processing constitutes a counterpart based on "visemes", the distinctive visual units of speech. Aiming at investigating the neural substrates of visemic processing in a disturbed environment, we carried out a simultaneous fMRI–EEG experiment based on discriminating syllabic minimal pairs involving three phonological contrasts, each bearing on a single phonetic feature characterised by different degrees of visual distinctiveness. The contrasts involved either labialisation of the vowels, or place of articulation or voicing of the consonants. Audiovisual consonant–vowel syllable pairs were presented either with a static facial configuration or with a dynamic display of articulatory movements related to speech production.

In the sound-disturbed MRI environment, the significant improvement of syllabic discrimination achieved in the dynamic audiovisual modality, compared to the static audiovisual modality was associated with activation of the occipito-temporal cortex (MT+V5) bilaterally, and of the left premotor cortex. While the former was activated in response to facial movements independently of their relation to speech, the latter was specifically activated by phonological discrimination. During fMRI, significant evoked potential responses to syllabic discrimination were recorded around 150 and 250 ms following the onset of the second stimulus of the pairs, whose amplitude was greater in the dynamic compared to the static audiovisual modality. Our results provide arguments for the involvement of the speech motor cortex in phonological discrimination, and suggest a multimodal representation of speech units.

1. Introduction

In speech perception, a stream of sounds produced by articulatory movements is decoded into a meaningful utterance. One prerequisite for understanding and producing an utterance is that speakers and listeners share the same linguistic code. The close relationship between speech perception and production is illustrated by difficulties experienced by congenitally deaf people struggling with oral language acquisition, and by impaired speech production in children with poor phonological awareness, who fail to acquire sufficiently well-defined speech representations required for reading acquisition (Peeters, Verhoeven, de Moor, & van Balkom, 2009).

Although, in everyday life, speech perception typically refers to verbal sound auditory processing, it is based on the integration of different sources of linguistic information, among which visual cues make a significant contribution. Indeed, depending on the speech-to-noise ratio, word length and verbal processing constraints, congruent facial display provides critical cues to healthy listeners in noisy environments (MacLeod & Summerfield, 1990; Sumby & Pollack, 1954), and benefits hearing-impaired listeners (Grant, Walden, & Seitz, 1998). Visual speech information can also alter the expected perceptual interpretation of clear auditory signals. The "McGurk effect", showing how adding incongruent visual information to an auditory speech signal can indeed change the final percept, provides strong support for a multimodal representation of speech (McGurk & MacDonald, 1976).

To specify the visual counterpart of a phoneme, Fisher (1968) coined the label of "viseme" to describe particular facial positions and movements that occur alongside speech production, and allow distinction between different configurations. Basically, if two phonemes can be distinguished through visual perception, without auditory cues, they belong to two different classes of visemes. Of course, several phonemes may fit into the same visemic class. For example in French, 17 classes of visemes comprise 36 phonemes (Le Goff & Benoît, 1996).

* Corresponding author at: Institut de Phonétique, Équipe de Recherche Parole et Cognition, U.R. 1339 – LILPA, Université de Strasbourg, 67084 Strasbourg cedex, France. Tel.: +33 3 68 85 65 64; fax: +33 3 68 85 65 69.
E-mail address: cdubois@unistra.fr (C. Dubois).

0028-3932/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2012.02.016

Explanatory speech constructs such as phonetic features, phonemes or visemes do not relate to discrete observable units. Whenever articulatory gestures are chained in a given order, they are co-articulated continuously through anticipation and perseveration effects, such that speech units actually greatly overlap. A classical view is to consider that continuous speech cues are mapped onto discrete units with reference to phonological representations, the characteristics of which are still a matter of debate (Galantucci, Fowler, & Turvey, 2006; Pisoni & Luce, 1987; Segui, 1984). Although there is no agreement about the specification of their unit, such mapping may be considered as context-dependent and multimodal. Hence, when mapping is mainly based on auditory signal it would consist in phonemic processing, and one may hypothesise that when visual cues are provided, visemes would be processed likewise. This will be referred to as "visemic processing".

Prior conceptions involving a clear-cut distinction between speech production and perception and their respective underlying neural substrates have been questioned both by neuropsychological investigations and functional neuroimaging studies. Over the last decades, among other theories of speech perception, several models referred also to speech production (Browman & Goldstein, 1990; Fowler, 1996; Liberman & Mattingly, 1985). Thus, the Motor Theory of Speech Perception made three major assumptions: first, that speech was processed by a specific module; second, that perceiving speech meant perceiving vocal tract gestures; and third, that the motor system was recruited for speech perception. Recent neuro-imaging studies found that simply listening to speech stimuli activated, in the frontal cortex, areas largely overlapping those activated when subjects actually produced similar speech stimuli (Wilson, Saygin, Sereno, & Iacoboni, 2004). These results support the view that speech perception implies an auditory-to-articulatory mapping process put forward by the Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

The recent discovery of the mirror neuron system, wherein neurons respond both when an action is performed by oneself or seen to be performed by someone else (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), supports knowledge about the central role of motor acts of speech production in speech perception. Rizzolatti and Arbib (1998) suggested that the mirror neuron system, which in human includes Broca's area and the premotor cortex, could support articulatory retrieval in the receiver of an oral message, arguing in favour of the view that speech gestures constitute the material of speech perception. The Perception-for-Action-Control Theory (PACT), proposed by Schwartz, Abry, Boë, and Cathiard (2002), also gained support from the existence of an action-observation matching system put forward by the discovery of the mirror neuron system. This theory considers that speech perception relies on a set of perceptual processes enabling, at a segmental level, to recover and specify the timing and targets of speech gestures, consisting of "perceptuo-motor" units (Schwartz, Basirat, Ménard, & Sato, in press). This framework is based on the assumption that action shapes perception and vice versa. As a result, "perceptuo-motor" units do not result from pure motor action, inasmuch as speech gestures also rely on perceptual – acoustic–auditory and visual – values, which in turn are useful to speech production.

The robustness of speech perception in spite of acoustic variability has been viewed as the result of multiple, complementary representations of the input based on both acoustic-phonetic and articulatory-gestural features (Scott & Johnsrude, 2003). This view received support from functional neuroimaging studies revealing that human speech perception might be based on multiple, hierarchical processing pathways recruited in the processing of different kinds of speech representations, involving among others acoustic-phonetic and articulatory-gestural aspects.

Despite some controversy about the early lateralisation of speech signal processing and its functional significance, current models of auditory speech processing agree that parallel streams are specialised in the analysis of different aspects of speech signals, following distinct processing pathways, dedicated to mapping sounds onto meaning and another involved in mapping speech sounds onto motor representations of articulation (Hickok & Poeppel, 2007; Obleser & Eisner, 2009; Rauschecker & Scott, 2009). The brain regions, which constitute these networks, encompass the superior temporal gyrus (STG) and middle temporal gyrus (MTG) including middle and/or posterior parts of the superior temporal sulcus (STS) bilaterally, as well as the premotor cortex or Broca's region.

Most cognitive models and phonetic theories of speech perception have been conceptualised around auditory speech processing and have not paid sufficient attention to visual cues inherent in conversational speech. The few fMRI studies having investigated visual and/or audiovisual (AV) speech perception looked at the possible involvement of the primary auditory cortex (PAC) in visual speech perception (Bernstein et al., 2002; Calvert et al., 1997; Pekkola et al., 2005), or of Broca's area in AV speech perception (Ojanen et al., 2005; Sekiyama, Kanno, Miura, & Sugita, 2003). The integration of AV cues in speech processing has been investigated using McGurk stimuli or audio and visual stimuli, during passive listening (Callan et al., 2003; Skipper, Nusbaum, & Small, 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007), in categorisation tasks (Jones & Callan, 2003) or in matching tasks (Saito et al., 2005). Speech discrimination tasks have been used to study phonological processing, specifically the ability to discriminate minimal phonemic contrasts in neuropsychological (Blumstein, Cooper, Zurif, & Caramazza, 1977) as well as fMRI investigations (Burton, Small, & Blumstein, 2000). These studies showed that the discrimination of phonemic contrasts involved the left frontal inferior gyrus in addition to the temporal cortex. Indeed, Blumstein et al. (1977) observed that aphasic patients with lesions involving both the left anterior and posterior cortices performed very poorly in phonemic discrimination tasks in comparison with aphasics with lesions localised only in the left temporal regions. In their neuroimaging study, Burton et al. (2000) showed that the discrimination of phonemic features recruited the left frontal gyrus in addition to bilateral activation of superior temporal gyrus (STG) when the discrimination task required further speech segmentation. Moreover, a recent study by Hutchison, Blumstein, and Myers (2008) found that activation in the right hemisphere increased as a function of the processing demand in a voice-onset time discrimination. As to the relevance of visual cues inherent in speech articulation to the discrimination of phonetic features, it has not been investigated so far.

In order to gain further understanding of the respective contribution of visemic and phonemic information to speech perception, we compared the effect of static and dynamic visual cues in a syllable pair discrimination task. Decision as to whether two syllables are similar or different requires the ability to distinguish minimal phonetic features, which can be performed upon auditory or visual speech perception. Whereas auditory cues may be sufficient to decide on the similarity of two syllables in a quiet environment, decision based on visual speech information alone is contingent on the articulatory properties of the syllables. Indeed, visual speech information is fragmentary and allows to perceive only some of the phonological contrasts. For example, speech-reading may in some instances be sufficient to discriminate the place of articulation, but not voicing or nasality. Moreover, discrimination is facilitated for anterior (labial or alveolar) *vs.* posterior (palatal or velar) places of articulation.

As the integration of visual information into speech processing significantly depends on the speech-to-noise ratio and on verbal

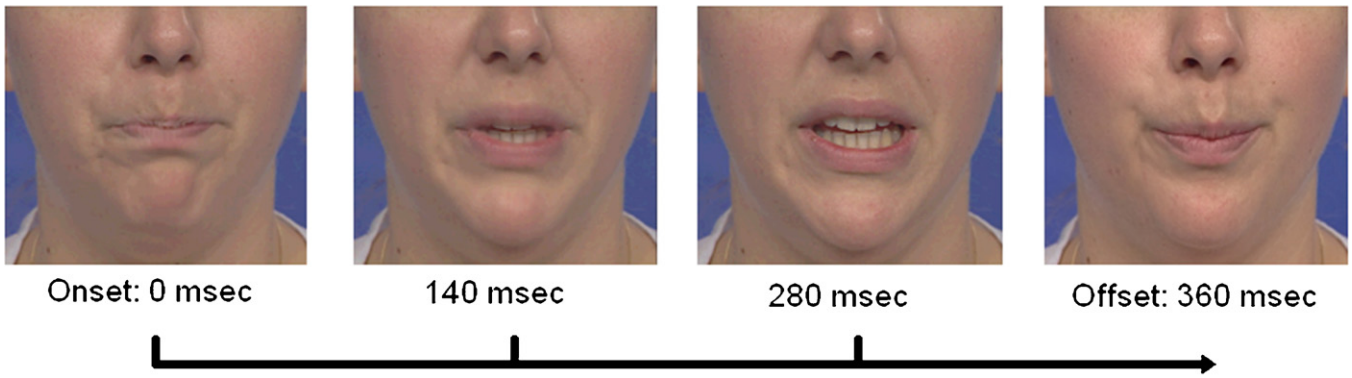

Fig. 1. Four frames of a dynamic AV stimulus corresponding to the start, the end and two midpoints of the articulation of the syllable /pi/.

processing constraints (MacLeod & Summerfield, 1990; Sumby & Pollack, 1954), we took advantage of the noisy environment inherent in fMRI scanning to investigate the neurophysiological basis of AV speech discrimination. To examine more specifically the influence of visual speech movements, brain responses to dynamic AV speech presentation were contrasted with responses to static AV speech presentation, and in order to look for a possible change in activation within the brain network recruited for AV speech processing, three minimal phonetic contrasts with different levels of visual distinctiveness were compared (labiality, place of articulation and voicing).

To complement event-related BOLD-contrast fMRI data with information about the time-course of dynamic AV speech processing, an electroencephalogram (EEG) was recorded simultaneously using interleaved EEG and fMRI measurements validated by Otzenberger, Gounot, and Foucher (2007). Several event-related potential (ERP) studies have investigated the neurophysiological correlates of AV integration in speech processing. They were mainly based on passive listening, as in Saint-Amour, De Sanctis, Molholm, Ritter, and Foxe (2007) who studied the mismatch negativity (MMN) component of the auditory ERP resulting from the automatic detection of incongruent AV speech units. Other studies involving identification tasks (van Wassenhove, Grant, & Poeppel, 2005) or recognition tasks (Besle, Fort, Delpuech, & Giard, 2004), showed decreased N1 latency and amplitude when processing congruent bimodal AV speech stimuli *vs.* unimodal auditory or visual stimuli. Such changes were found in target detection or forced-choice procedures performed in a quiet environment, which allows to focus on auditory ERPs *per se*.

However, studying AV integration in speech processing in the comparatively disturbed environment of an EEG/fMRI set-up involves experimental conditions that differ in several respects from those encountered in typical behavioural or neurophysiological research. For one thing, the timing of the task should be relevant for both the fMRI and electrophysiological investigations. In addition, syllable discrimination, made more difficult by the scanner noise, may rely to a larger degree on the visual component than it might otherwise. Consequently, one may predict that dynamic visual cues should, to a degree, facilitate phonological discrimination and influence the latency and amplitude of the ERPs evoked by the discrimination of minimal phonetic features in such noisy background. In view of our unusual paradigm, *i.e.* syllable discrimination, and of the different experimental context, simultaneous EEG/fMRI, we therefore did not expect to replicate the results of previous electrophysiological studies (Besle et al., 2004; van Wassenhove et al., 2005) but rather anticipated that facilitation of minimal phonetic feature discrimination in the noisy environment, mediated by dynamic visual cues would result in electrophysiological changes in the ERPs, varying as a function of the degree of visual distinctiveness of the features to be discriminated.

With respect to fMRI, we predicted increased activation in components constituting the network dedicated to auditory speech discrimination and, potentially involved in "visemic processing". The PAC and STG, even the premotor cortex might be involved in this processing, as might the Sylvian parieto-temporal (Spt) area and the mid-posterior STS, considered respectively as sensory-motor and AV integration sites (Hickok & Poeppel, 2004; Scott & Wise, 2004). In addition, when subjects were provided with dynamic visual cues to perform phonological discrimination tasks, we expected bilateral occipito-temporal activation (area MT + V5) related to visual movement processing.

2. Materials and methods

2.1. Subjects

The subjects were 30 adult French native speakers (mean age = 22.6; range = 18–27, 15 females), who reported no neurological or psychiatric deficit, with normal hearing and vision (or corrected to normal), and had not experienced any verbal learning disability. All subjects were recruited among students of the University of Strasbourg, and were right-handed according to the Edinburgh handedness inventory (Oldfield, 1971). The volunteers first acquainted themselves with the MRI scanner environment by lying down within the scanner for a few minutes, and were then made aware of the purpose and conditions of the experiment, namely that they would be wearing a set of electrodes (in a magnetic environment). All participants gave their informed consent prior to their inclusion in the study, and were paid for their participation in the investigation. The local ethical committee (CCPPRB Alsace no. 01) approved the experimental procedure.

2.2. Stimuli

In the dynamic AV presentation modality, the stimuli consisted in a set of eight French syllables and eight non-phonological stimuli. The eight consonant–vowel (CV) syllables [pi] [py] [bi] [by] [ti] [ty] [di] [dy] were produced by a female native speaker of French and recorded in video files lasting for 360 ms each. Each video showed the bottom part of the speaker's moving face (frame rate 25 images/s, audio sample rate 44.1 kHz in 16 bits) starting and ending with a closed mouth position (Fig. 1). The non-phonological stimuli consisted in videos of natural syllables [fa] [fo] [va] [vo] played backward. Additionally, in order to obliterate any phonological cues, their audio part was altered using Audacity Software (http://audacity.sourceforge.net/) by applying a low-frequency oscillation effect (5 Hz, start phase at 180°, deepness 75%, resonance factor 6, and periodicity 85%). To ensure that these stimuli were non-phonological, they were presented in an identification task to 55 subjects who were asked to write down the perceived syllables. In this task, 83% of these non-phonological stimuli could not be identified as French syllables.

In the static AV modality, the same set of stimuli was used, but the moving face was replaced with a stilled face, randomly selected out of the video frames of the stimuli. In each pair, the two audio stimuli were presented with the same still.

In a preliminary study, the visual distinctiveness of the syllables to be displayed in the dynamic AV mode in the fMRI protocol was evaluated in a forced-choice task, administered to a group of 17 subjects not involved in the subsequent fMRI experiment (mean age = 21.7; range = 18–34, 10 females). In this task, the subjects, facing a screen, were presented with a silent video of three syllables (*e.g.* [pi pi py]) and asked to decide, by clicking on a two-button mouse, which of the second (left

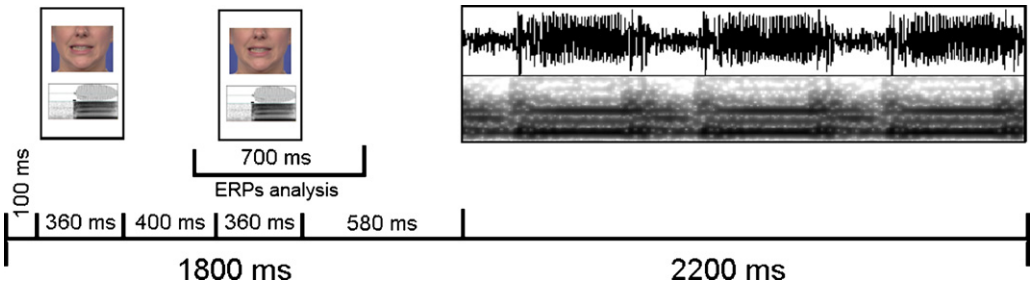

Fig. 2. Timing of interleaved EEG/fMRI design of the stimuli presentation during a complete trial including gradients. Scheme of a complete trial as set up for interleaved fMRI acquisition and EEG recording.

button) or third (right button) syllable was identical to the first of the triplet. The results showed that the phonetic feature of vocalic labiality (*i.e.* [bi by bi]) and place of articulation (*i.e.* [pi pi ti]) were significantly ($F(3,48) = 40.6$, $p < 0.000001$) more efficient visual cues than voicing (*i.e.* [bi pi bi]), which was then considered as a visually neutral feature, since voiced and voiceless syllables cannot be distinguished visually.

2.3. Experimental procedures

The task consisted in forced-choice discrimination between pairs of different and similar AV stimuli. The trials lasting for 1120 ms consisted of two 360-ms AV stimuli displayed on a screen, with a 400-ms inter-stimulus interval (ISI) (plain blue screen) between them. Each trial was preceded by a 100 ms pre-stimulus period which was also followed by a period of 580 ms in order to minimise the impact of gradients on EEG recordings (Fig. 2). The subjects perceived the AV stimuli through prismatic goggles and headphones with high noise attenuation (AVOTEC SS-3100 silent scan). They did not wear earplugs in order to preserve the auditory perception. After each trial they had to decide whether the two stimuli were similar or different by clicking on one of a two-button mouse. In each dynamic or static modality, the participants were presented with a total of 160 trials with pairs of different stimuli (40 of each category; one set of non-phonological and three of phonological pairs), besides 40 empty trials which served as null events. The trials were pseudo-randomly distributed using Optseq (http://surfer.nmr.mgh.harvard.edu/optseq) which automatically optimises the order and timing of events for event-related fMRI experiments (Dale, 1999), on the basis of the duration and sampling time of the haemodynamic responses, the number and types of stimuli and null events. In each modality of presentation, the task was performed within 24 min (360 4-s trials), the complete acquisition lasted for 48 min.

The three sets of phonological pairs were designed to examine the influence of visual speech cues relevant to three classes of phonetic features, categorised as corresponding to three different visemes. The first contrast between the spread [i] *vs.* rounded [y] vowels, was based on the feature of labiality. The second contrast was relevant to the place of articulation, *i.e.* labial [p b] *vs.* coronal consonants [t d] and the last, voicing-based contrast, opposed voiced [b d] *vs.* voiceless consonants [p t]. We will refer to these contrasts as "labiality", "place of articulation" and "voicing". Accuracy and response times (RT) measured from the onset of the second syllable in each trial were recorded.

2.4. MRI acquisition

MR images were acquired using a Bruker 2T S200 scanner (Bruker Medical GmbH, Ettlingen, Germany), equipped with an SK330 insert gradient coil (30 mT/m, 150 μs rise time) and a radio-frequency head coil. BOLD-contrast functional images were obtained using echo planar imaging (EPI) and an intercommissural AC-PC slice orientation (32 slices, matrix size = 64 × 64 pixels, voxel size = 4 × 4 × 4 mm, TE = 43 ms) single shot following an event-related design. The repetition time was set to 4 sec, encompassing a 1.8-s gradient artefact-free window for interleaved EEG recording. The anatomical images were acquired with a fast spin echo sequence (T2-weighted) and a resolution of 2 × 2 × 2 mm.

After discarding the first three volumes used to reach signal equilibrium in each modality of presentation, a total of 360 volumes (320 volumes with presentation of pairs of similar or different stimuli and 40 volumes without stimuli) were acquired.

2.5. EEG signal acquisition

EEG signals were acquired continuously during fMRI acquisition using a magnetic resonance-compatible system (EMR32: Schwarzer, Munich, Germany), equipped with a digital signal processor board, which also received three synchronised trigger inputs: from the stimuli, the MRI volumes' onset dating, and a separately amplified electrocardiogram ECG channel (Physiogard, Bruker, SARL, Wissembourg, France). The ERPs were analysed in a 700 ms window encompassing 100 ms before the onset of the second stimulus and 600 ms post onset (Fig. 2).

A set of 23 Ag/AgCl electrodes, with iron-free copper leads fixed individually using EEG paste (Elefix, Nikon Khoden) served to record EEG data and eye movements. EEG signals were recorded from 19 electrodes positioned according to the international 10-20 system (F7, F3 Fz, F4, F8, FC3, FC4, T3, C3, Cz, C4, T4, CP5, CP6, T5, P3, Pz, P4, T6) with nasion reference, and the three remaining electrodes monitored horizontal and vertical eye movements, for subsequent off-line validation (*i.e.* rejection of artefact-ridden recordings). Channel impedance tested outside the magnet was kept below 5 kΩ for each electrode. The signals were sampled at 1 kHz and filtered from 0.05 to 70 Hz. They were recorded and displayed using the Brainlab software (OSG, Rumst, Belgium).

2.6. Data analysis

2.6.1. Behavioural data analysis

Statistical analyses performed on the accuracy rate and response times consisted of repeated-measures analyses of variance, with the two modalities of presentation (static and dynamic AV) and four contrasts (3 phonological: labiality, place of articulation, voicing and 1 non-phonological) as factors, and were followed by post hoc analyses (Least Significant Difference, $p < 0.05$). Since in the static modality 10 participants performed under chance in discriminating the feature of labiality, they were not included in the analyses.

2.6.2. fMRI data analysis

The fMRI data were analysed using Statistical Parametric Mapping 5 software (SPM5, Wellcome Department of Imaging Neuroscience, London, UK; Friston, Ashburner, Kiebel, Nichols, & Penny, 2007) implemented in Matlab V6.1 (Mathworks, Sherborn, MA, USA). Four subjects were discarded owing to their inability to perform the entire experiment. The EPI images were corrected for motion and repositioning, spatially normalised into standard co-ordinates based on the Montreal Neurological Institute (MNI) reference brain, and smoothed spatially with an 8-mm full-width half-maximum (FWHM) Gaussian kernel. Low frequency drifts were removed using a high-pass temporal filter. Only images corresponding to correct responses delivered by the subjects were taken into account in this analysis. Differences between task and baseline activation were assessed using the general linear model, yielding *t*-statistics for each voxel. In a first-level analysis, we calculated, for each participant, in each modality of presentation, the contrast images for the three phonological and one non phonological control items. The resulting images were entered into a group random-effect analysis, allowing generalisation to the whole population. All reported areas of activation were significant using a $p < 0.01$ corrected (Family Wise Error, FWE), with a spatial extent threshold of 25 voxels. The anatomical localisation of the local maxima and clusters was assessed with reference to the MNI space.

The whole-brain analysis was intended to identify the brain network activated by the discrimination of syllables, and to study the influence of visual cues, *per se*, by comparing the two modalities of speech presentation. To investigate neural processes in brain regions theoretically relevant to visual and auditory speech processing, we delineated using the MarsBar toolbox (Brett, Anton, Valabregue, & Poline, 2002) nine regions of interest (ROIs). All ROIs were 5-mm spheres except in the PAC and Broca's region (6-mm spheres). For each ROI, the mean contrast was computed in each subject using the Marsbar toolbox. Activation within a ROI was considered consistent when it was significant in more than 50% of the individuals. Then the individual results were entered in an analysis of variance, with stimuli and modality of presentation as factors, to test the effects of the phonological contrasts and of the static and dynamic AV display modes. In reference to studies showing the involvement of Broca's area in audio and visual speech processing, we placed a 6-mm ROI centred on the mean coordinates ($x = -46$; $y = 17$; $z = 11$) calculated from Ojanen et al. (2005) and Sekiyama et al. (2003) studies. As activation within this ROI was found significant only in 39.4% of the individuals, it was not further analysed.

Two ROIs were centred bilaterally in the left and right PAC on the mean coordinates ($x = -46$; $y = -23$; $z = 4$ and $x = 45$; $y = -22$; $z = 5$) of the activation peaks obtained for left and right monaural pure tone stimulation in Devlin et al. (2003). Two spheres, in the lateral temporo-occipital regions tracking visual movement were centred on the coordinates ($x = -46$; $y = -77$; $z = -7$ and $x = 43$; $y = -71$; $z = -6$)

corresponding to Campbell et al.'s findings about visual speech *vs.* stilled face contrast (Campbell et al., 2001). Two ROIs, in the mid-posterior STS, were centred on coordinates ($x=-57$; $y=-31$; $z=1$ and $x=53$; $y=-29$; $z=1$) on the basis of data from a meta-analysis of neuroimaging studies on sublexical speech perception (Turkeltaub & Branch Coslett, 2010). The centre of the ROI around the Spt area in the left hemisphere was made to coincide with the mean coordinates ($x=-57$; $y=-36$; $z=16$) of activation peaks obtained in five studies aiming at identifying human auditory regions with both sensory and motor response properties (Buchsbaum et al., 2005; Hickok, Buchsbaum, Humphries, & Muftuler, 2003; Okada & Hickok, 2006; Pa & Hickok, 2008). The ROI situated in the premotor cortex was centred on the mean coordinates ($x=-51$; $y=-6$; $z=49$) of activation peaks found in studies on motor speech perception during speech listening (Fridriksson et al., 2008; Okada & Hickok, 2009; Wilson, Molnar-Szakacs, & Iacoboni, 2008; Wilson et al., 2004).

2.6.3. ERP data analysis

The 1.8-s gradient artefact-free windows within the simultaneous fMRI-EEG recording allowed the analysis of a subset of EEG data for ERP computation. The interleaved acquisition mode allowed taking advantage of the delay between the haemodynamic response and the electrical activity. EEG data were then analysed and computed using Matlab 6.1 (Mathworks). According to the method proposed by Otzenberger et al. (2007), the cardio-ballistic artefact within these windows was removed for each subject, by regressing a modelled pulse artefact using the ECG as a guide. To this purpose a sample of ECG signal was recorded for each subject before starting the fMRI acquisition. Low- and high-frequency components of the signals were removed using a Fourier band-pass filter of 0.1–30 Hz.

The EEG data of only 11 subjects were included in the ERP analysis, owing to frequent signal artefacts and eye movements in the others. After signal artefact and eye movement correction, the ERPs evoked in response to correct phonological discrimination were then computed from 38 ± 2 trials for each contrast. Analyses were carried out on epochs ranging from −100 ms to 600 ms relative to the onset of the second AV stimulus of each trial, using Statistical Parametric Mapping Software (SPM5, Wellcome Department of Imaging Neuroscience, London, UK; Friston et al., 2007), a mass univariate approach in which spatiotemporal data are modelled within the statistical framework of the general linear model (Kiebel & Friston, 2004a, 2004b). The method requires no a priori subjective definition of response peaks of interest within the ERP signal, and allows simultaneous comparisons across time points within predefined time-windows and across EEG channels (Myatchin, Mennes, Wouters, Stiers, & Lagae, 2009). In the pre-processing analysis, the average ERP signal of each channel was interpolated on the scalp surface, resulting in a 2-D image time series for each subject, in each modality (static and dynamic AV) and for each of the three different sets of phonological pairs to be discriminated. In the first-level analysis, time was entered as the second dimension. We explored the ERP data over five 50-s windows centred on 100, 150, 200, 250, 300 ms following the onset of the second stimulus in each pair. In the second-level group analysis we first tested the local maxima of the amplitude for each modality and each phonological contrast and in a two-sample *t*-test the effect of modality on each phonological contrast. The significance threshold was set at $p<0.01$ for multiple comparisons. In each 1800 ms gradient-free period, the 700 ms window was excerpted for ERP analysis, 760 ms after the gradient of the previous trial, and finishing 340 ms before the next gradient.

3. Results

3.1. Behavioural results

The participants completed the discrimination tasks with an overall rate of accuracy above 90%, except for the syllable pairs contrasting in respect of labiality ([i] *vs.* [y]) in the static AV modality (Table 1).

The ANOVA on the rate of accuracy revealed a significant main effect of AV modality ($F(1,15)=11.56$, $p<0.003$) and contrasts ($F(3,45)=9.74$, $p<0.00004$) with a significant interaction ($F(3,45)=11.30$, $p<0.0001$). Post hoc analyses showed significantly better discrimination in the dynamic AV presentation mode (96.8%) compared to the static AV display mode (92.1%). With respect to phonological contrasts, they showed a significantly greater accuracy in response to contrasts opposing the place of articulation (96.8%) or voicing (97.2%) in comparison to labiality ([i] *vs.* [y]) (90.2%), which did not differ significantly from the performance achieved for the non phonological contrast (93.4%). The interaction was related to less accurate discrimination of labiality in the static AV modality (82.8%) compared to all other contrasts.

Regarding response times, only a main effect of phonological contrasts ($F(3,45)=9.1$, $p<0.00007$) was found with a significant interaction ($F(3,45)=3.05$, $p<0.04$). Post hoc analyses revealed that non-phonological pairs were discriminated faster than phonological pairs, with no significant difference as a function of the contrast-related set these pairs belonged to (Table 1).

Table 1
Average and standard deviation of accuracy rate and response time (RT), for the discrimination of the four contrasts, in dynamic and static AV presentation.

AV modality/contrast	Accuracy (%)	RT (ms)
Labiality		
Static	82.8 ± 11.3	827 ± 147
Dynamic	97.7 ± 2.9	772 ± 178
Place of articulation		
Static	95.8 ± 4.1	805 ± 191
Dynamic	97.8 ± 3.9	766 ± 154
Voicing		
Static	97.5 ± 2.2	777 ± 137
Dynamic	96.9 ± 5.3	788 ± 174
Non phonological		
Static	92.3 ± 5.3	734 ± 144
Dynamic	95 ± 8.4	746 ± 208

3.2. fMRI results

3.2.1. Whole-brain analyses

The statistical maps showing significantly increased activation relative to rest during discrimination of phonological contrast-related pairs, and non phonological pairs in static and dynamic stimulus presentation are shown in Fig. 3.

In the dynamic AV presentation mode, correct discrimination of phonological as well as non phonological pairs activated a neural network involving bilaterally the STG, MTG (BA 22/41 and 21) and the occipito-temporal region (BA 19 and 37) in the right hemisphere. Additional activation in the left premotor cortex (BA 6) was observed as subjects were engaged in the discrimination of phonological pairs contrasting in respect of voicing or place of articulation (Fig. 3 and Table 2).

In the static AV presentation mode, significant activation involved only the STG and MTG bilaterally.

The dynamic AV > static AV contrast (Table 3) revealed significantly greater activation in the occipito-temporal region (BA 19/37) in the right hemisphere no matter whether discrimination involved phonological or non phonological pairs. Only the discrimination of phonetic opposition based on vocalic labiality elicited symmetrical activation, but less extended in the left hemisphere.

3.2.2. ROI analyses

The analysis of variance with pairs and modalities of presentation as factors showed a significant main effect of pairs on the mean contrasts computed in four ROIs, in the left ($F(3,75)=2.81$, $p<0.045$) and right PAC ($F(3,75)=3.71$, $p<0.015$), the left Spt ($F(3,75)=4.76$, $p<0.004$) and mid-posterior STS, bilaterally. Post hoc analysis showed that activation in the PAC and the Spt was significantly higher in response to the discrimination of syllable pairs opposing voicing compared to those opposing vocalic labiality features, or the place of articulation. Bilateral activation in the mid-posterior STS was strongest in response to the discrimination of the place of articulation, relative to the other contrasts, whether phonological or not.

Analysis of the ROI in the left premotor cortex showed a significant main effect of pairs ($F(3,75)=5.24$, $p<0.002$) related to significantly higher activation elicited by the discrimination of phonological *vs.* non phonological pairs. A significant main effect of the modality of presentation was found only for bilateral activation in the occipito-temporal region, which was enhanced by dynamic

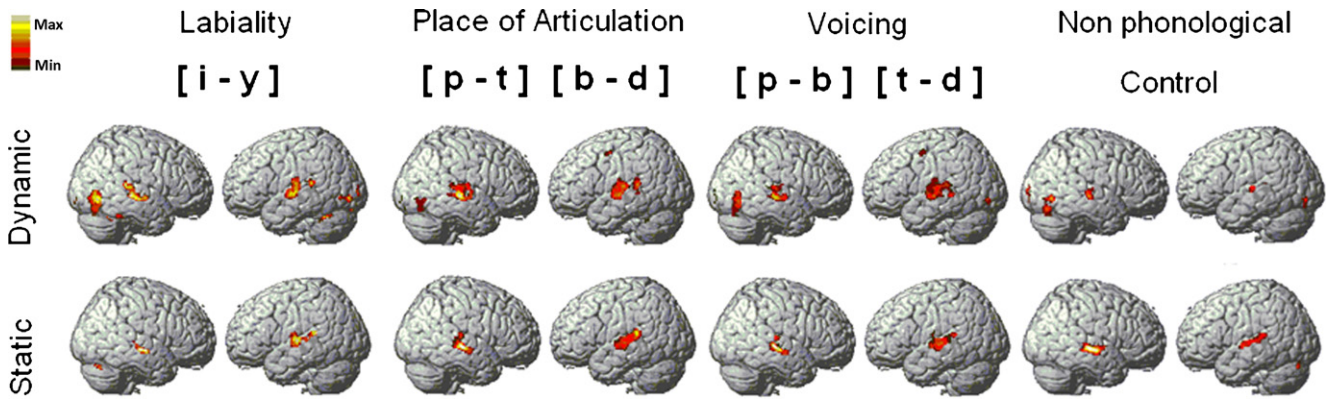

Fig. 3. Activation areas to correct discrimination of the three phonological contrasts, in the dynamic (top) and static (bottom) AV presentation. Activation significant at p FEW-corrected < 0.005, with an extent threshold of 25 voxels.

Table 2
Significant activation elicited by discrimination of the four contrasts in dynamic and static AV presentation.

Brain regions	BA	Static presentation					Dynamic presentation				
		Voxels	x	y	z	T	Voxels	x	y	z	T
Labiality											
L. superior temporal gyrus	22/41	179	−58	−28	6	8.35	160	−58	−40	14	9.68
								−38	−30	10	9.57
								−46	−38	10	8.01
	22	57	−62	−38	14	9.02	324	−60	−10	0	10.03
L. inferior occipital gyrus	18/19						74	−40	−86	−6	10.84
								−42	−74	−4	8.61
R. middle temporal gyrus	21	106	58	−20	−8	9.42	534	64	−20	−6	10.71
			66	−12	−8	8.57		64	−36	−10	9.59
		144	44	−32	6	9.02		60	−12	−6	9.40
R. inferior temporal gyrus	37/19						367	48	−68	−4	9.68
								46	−76	−4	9.31
								42	−74	−20	8.67
Voicing											
L. superior/middle temp. gyrus	22	297	−62	−20	4	9.67	442	−64	−38	0	9.63
			−62	−10	0	8.19		−62	−32	12	9.22
			−62	−40	14	8.47					
L. superior temporal gyrus	41						47	−42	−40	10	8.50
								−36	−36	6	7.88
L. premotor cortex	6						27	−50	−2	52	9.91
R. superior temporal gyrus	22/41	388	66	−26	2	10.58	409	64	−22	−4	12.79
			46	−22	2	9.79		64	−20	6	8.66
			44	−30	6	9.09		44	−30	6	8.27
R. inferior temporal gyrus	37/19						261	48	−68	−4	8.36
								44	−74	−20	8.35
Place of articulation											
L. superior temporal gyrus	22	463	−64	−20	4	11.08	428	−64	−18	4	11.28
			−64	−32	10	10.07		−62	−8	4	10.62
	41		−52	−34	14	8.76	111	−48	−38	18	8.89
								−38	−32	12	8.45
L. middle temporal gyrus	21						98	−60	−38	10	10.13
								−60	−40	0	7.63
L. premotor cortex	6						25	−50	−4	52	10.59
R superior/middle temporal gyrus	21/41/42	508	64	−22	−4	12.27	742	64	−24	−2	14.24
			52	−26	−2	9.81		64	−16	10	9.78
			44	−32	6	9.09					
R. inferior occipital gyrus	17/19						189	10	−94	−6	9.90
	19						91	48	−74	−12	8.72
Non phonological (control)											
L. superior temporal gyrus	41/22	45	−36	−30	8	9.15	41	−44	−38	12	8.23
		222	−64	−20	6	10.13					
			−60	−10	0	9.87					
			−62	−36	14	9.41					
R. middle temporal gyrus	21/22	463	62	−22	−4	10.60	97	66	−24	0	9.56
			62	−14	−6	12.22					
			46	−24	2	10.10					
R. inferior occipital/temporal gyrus	19/37						136	42	−74	−20	9.21
								48	−68	−6	8.16

All reported cerebral regions are significant at p FEW-corrected < 0.005, with an extent threshold of 25 voxels. The coordinates are provided in the MNI template. BA: Brodmann's areas. Coordinates (x, y, z) are those of the local maxima of the cluster expressed according to the Montreal Neurological Institute Standard Brain (MNI system).

Table 3
Brain areas in which dynamic AV presentation induced significantly higher activation relative to static AV presentation.

Dynamic AV > Static AV contrast	Brain regions	BA	Voxels	*x*	*y*	*z*	*T*
Labiality	L. occipito-temporal	19	86	−44	−68	2	4.75
	R. occipito-temporal	19/37	509	42	−66	−6	6.33
Place of articulation	R. occipito-temporal	19/37	140	48	−66	−6	4.83
Voicing	R. occipito-temporal	19/37	342	42	−66	−6	5.79
Non phonological	R. occipito-temporal	19/37	295	48	−68	−6	5.71

All cerebral regions reported are significant using $p < 0.001$ uncorrected at the voxel level. BA: Brodmann's areas. Coordinates (x, y, z) are those of the local maxima of the cluster expressed according to the Montreal Neurological Institute Standard Brain (MNI system).

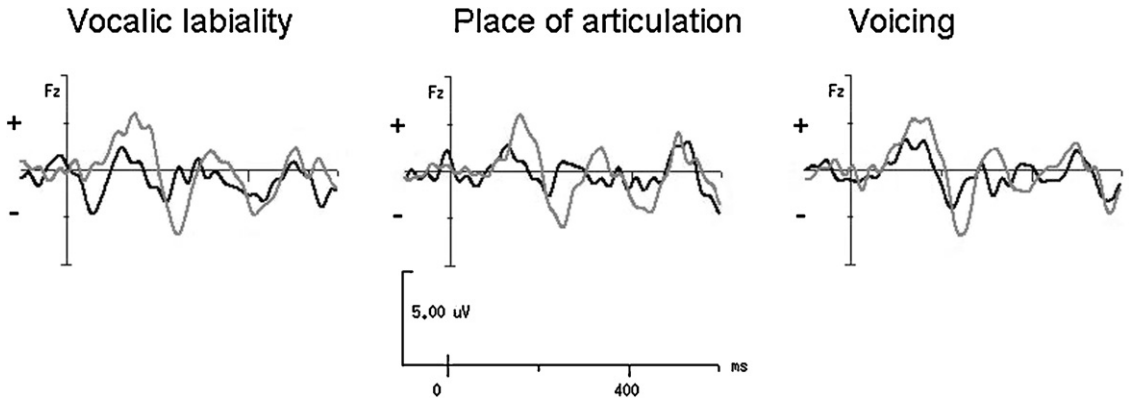

Fig. 4. ERPs to correct discrimination of the three phonological contrasts, in the dynamic (grey line) and static (black line) AV presentation. The time window is from −100 ms to 600 ms, and the vertical line indicates the onset of the second item of contrasting pair.

AV displays compared to static AV displays (Left: $F(1,25) = 19.04$, $p < 0.0002$; Right: $F(1,25) = 50.05$, $p < 10^{-6}$).

3.3. ERP results

Fig. 4 shows the grand-average ERP waveforms from the Fz electrode site, over trials in which correct discrimination of syllable pairs was achieved, for each phonological contrast, and each modality of presentation. These waveforms were displayed with the ELAN software pack for electrophysiological analysis, developed at INSERM U821, France.

In the dynamic AV modality, correct discrimination of syllable pairs elicited significant positive waves within the 125–175 ms window, for the 3 phonological contrasts, as follows: labiality: ($T = 6.32$, $p < 10^{-4}$; $T = 4.75$, $p < 10^{-4}$); place of articulation: ($T = 6.81$, $p < 10^{-4}$; $T = 6.61$, $p < 10^{-4}$); voicing: ($T = 4.97$, $p < 0.001$; $T = 4.14$, $p < 10^{-4}$), at Fz and Pz electrodes, respectively. Significant negative waves were recorded within the 225–275 ms window at Fz and Cz electrodes when discriminating the labiality ($T = 6.42$, $p < 10^{-4}$), place of articulation ($T = 5.74$, $p < 10^{-4}$) and voicing contrasts ($T = 3.96$, $p < 0.001$).

In the static AV modality, correct discrimination of syllable pairs elicited a significant positive wave at the Fz electrode in the 125–175 ms window only for the voicing-based contrast ($T = 4.57$, $p < 0.01$). Within the 225–275 ms window, significant negative waves were recorded at different electrode sites for the different phonological contrasts, as follows: at Fz and Cz electrodes for voicing ($T = 4.71$, $p < 10^{-4}$); at Cz for the place of articulation ($T = 3.19$, $p < 10^{-4}$) and at CP5 for labiality ($T = 3.29$, $p < 0.0002$).

In the 125–175 ms window the difference in amplitude as a function of the modality of stimulus presentation was only significant for one contrast, *i.e.* labiality, which elicited waves of higher amplitude at the Fz ($T = 3.17$, $p < 0.004$) and Pz ($T = 3.69$, $p < 10^{-4}$) electrodes in the dynamic *vs.* static AV modality.

As to the 225–275 ms window, the difference in negative ERP amplitudes as a function of the modality of stimulus presentation was significant for all three contrasts, which, in the dynamic AV mode, elicited at the Fz electrode waveforms of higher amplitudes as follows: labiality: $T = 4.67$, $p < 10^{-4}$; voicing: $T = 2.95$, $p < 0.009$; and place of articulation: $T = 3.37$, $p < 0.008$. The latter additionally elicited a negative waveform of increased amplitude at Pz ($T = 2.92$, $p < 10^{-4}$).

No significant change in latency was found, except where discrimination involved the place of articulation, in the 225–275 ms window, in which negative potentials were evoked significantly earlier at the Fz electrode ($F(1,10) = 9.5$, $p < 0.01$), when the stimuli were presented in the static AV modality.

4. Discussion

The aim of the present study was to investigate the neural substrates underlying the visemic dimension of speech processing in the sound-disturbed environment of an EEG-fMRI experiment.

In this environment, the contrast opposing vocalic labiality features was less accurately discriminated than contrasts opposing voicing and place of articulation features, for which subjects scored at ceiling. This may be explained by the physical characteristics of the fMRI scanner noise, which may have affected the perception of the third formant known to be critical for the identification of vowels with contrasted labiality features (Abry & Boë, 1986). This contrast was also the only one, which significantly benefited from the dynamic visual cues added to auditory signal. This suggests, first that better understanding of words and sentences typically reported when visual speech cues are provided in a noisy environment (MacLeod & Summerfield, 1990), results from improved speech processing rather than from resorting to higher-level verbal processing. Secondly, pointing to the involvement of visemic processing, *i.e.* the processing of facial positions and movements related to speech production, in the improvement of basic phonological processing this result also supports the claim about the existence of perceptuo-motor representations put forward by the Perception-for-Action-Control Theory (Schwartz et al., in press).

In accordance with current functional anatomic models of speech processing, in the present study the discrimination of phonetic features consistently activated bilaterally the STG and MTG, including the mid-posterior STS, in dynamic as well as static AV presentation modes. However, when the stimuli were presented dynamically, the discrimination task activated additional brain regions, *i.e.* the right occipito-temporal cortex (BA 19/37) recruited for all pairs of stimuli, and the left premotor cortex (BA 6) for pairs contrasting in respect of voicing or place of articulation. Moreover, as shown by electrophysiological data, the cues related to articulatory speech configuration provided by dynamic AV presentation induced significant changes in evoked potentials around 150 ms and 250 ms after the onset of the second syllable of each pair.

4.1. Influence of dynamic visual cues on the speech perception network

The whole-brain analysis revealed that the brain region, significantly more activated when subjects discriminated the four different pairs of audio visual stimuli displayed dynamically as opposed to statically, was localised in the right occipito-temporal cortex (Table 3) at the boundary of BA 19 and BA 37 ($x = 42–48$; $y = -66$ to -68; $z = -6$), overlapping the area delineated by (Hasnain, Fox, & Woldorff, 1998) as the functional area MT + V5, involved in visual motion processing ($x = 42$; $y = -67$; $z = -6$). In line with studies showing activation of this area by various kinds of movements (for a review see Bartels, Logothetis, & Moutoussis, 2008), including speech- and non-speech-related facial movements (Calvert & Campbell, 2003; Hall, Fussell, & Summerfield, 2005), its involvement in our study was not specifically related to the discrimination of phonological features. As to the left-lateralised MT + V5, it was activated only by phonological discrimination involving vocalic labiality. With respect to this area, this right–left difference is actually consistent with findings denoting right hemispheric dominance for visual attention (Marshall & Fink, 2001) and higher inter-subject functional and anatomical variability of the left MT + V5 (Wilms et al., 2005). However, the ROI analysis revealed a significant bilateral increase in activation in MT + V5 when gestural information was available to perform the discrimination tasks.

Contrary to our expectation, visemic processing, brought into play when visual cues related to speech articulation were available, did not significantly increase activation in brain areas known to be involved in either phonemic processing, *i.e.* the PAC, or integrative processing, *i.e.* the mid-posterior STS.

The PAC, involved in silent visual speech (Calvert et al., 1997; Pekkola et al., 2005), was, in our study, activated by phonological discrimination, regardless of the nature of the visual information provided to the subjects, and so was the left-lateralised Spt, to which sensory-motor integration is imputed (Hickok & Poeppel, 2007). The fact that, in the present study, the subjects were provided with visual cues, even minimal as was the case in the static modality of presentation, in addition to auditory information may explain the lack of significant increase in activation in these areas.

With respect to the left mid-posterior STS, its activation in the discrimination of both phonological and non-phonological AV stimuli, regardless of the modality of presentation, is consistent with its implication in the integration of different perceptual inputs irrespective of stimulus category, such as objects (Beauchamp, Lee, Argall, & Martin, 2004), typographic characters (van Atteveldt, Formisano, Goebel, & Blomert, 2004), iconic gestures (Holle, Obleser, Rueschemeyer, & Gunter, 2010) or AV speech (Calvert & Campbell, 2003; Campbell et al., 2001). Several neuroimaging studies have shown that the involvement of the left STS in AV speech processing, depended on the degree of AV mismatch (Jones & Callan, 2003). Activation elicited by matching AV speech stimuli was found to increase in response to stimuli with conflicting auditory and visual components (Ojanen et al., 2005). A recent MEG study of audiovisual speech error prediction evidenced different spatial distribution of cortical activity in response to incongruent *vs.* congruent AV speech stimuli (Arnal, Wyart, & Giraud, 2011). Specifically, they showed that when auditory signals invalidated predictions inferred from visual percept, oscillatory responses, scaled with the degree of audio–visual congruence, were elicited locally in the left STS area. Finally, Callan et al. (2003) and Sekiyama et al. (2003) observed a significant increase in the left posterior STS activation during AV speech processing under acoustic noise and low intelligibility conditions. The absence of significant increase in brain activation in response to dynamic *vs.* static AV display of syllables in our study may be explained by the fact that, in both conditions, the subjects were presented with congruent audio–visual speech only, in an consistently disturbed environment.

4.2. Involvement of the left premotor cortex in visemic processing

The whole-brain analysis showed significant activation in the left premotor cortex resulting from the discrimination of features relating to voicing or the place of articulation, when pairs of stimuli were presented in the dynamic AV modality.

The ROI analyses revealed that discrimination of either of the three phonological contrasts activated the left premotor cortex significantly more than discrimination of non-phonological pairs of AV stimuli. The implication of the premotor cortex in audiovisual speech processing has been observed in most brain imaging studies (Callan et al., 2003; Campbell, 2008; Ojanen et al., 2005; Pekkola et al., 2005; Skipper et al., 2005; Watkins, Strafella, & Paus, 2003; Wilson et al., 2004). In many, it was activated together with Broca's area, the activation of which however appeared to depend on audiovisual mismatch or stimulus ambiguity (Ojanen et al., 2005; Pekkola et al., 2005; Skipper et al., 2007; Wilson et al., 2004).

In the present study, the level of activation did not vary as a function of the phonological contrast involved, or as a function of stimulus presentation modality. The fact that it was not recruited for non-phonological stimulus processing is consistent with findings by Fridriksson et al. (2008), who observed higher activation in the left premotor cortex in discriminating silent speech movements in contrast to non-speech movements, which activated preferentially the parietal region. Such contrast in the involvement of the left premotor cortex in processing speech *vs.* non-speech movements, also found in earlier studies (Campbell et al., 2001; Hall et al., 2005), suggests that the processing of speech movements is special and distinguishable from non-speech movements.

One should mention that the peak of activation found in our AV speech discrimination tasks ($x = -50$; $y = -2/-4$; $z = 50$) is very close to that found in this area for purely visual speech ($x = -51$; $y = -7$; $z = 54$) in Fridriksson et al. (2008), and that reported by Wilson et al. (2004) ($x = -50$; $y = -6$; $z = 47$) which was found to be common to the passive perception of CV syllables and to the actual production of these same syllables. Wilson et al. (2004) considered this result as an argument in favour of an auditory-to-articulatory mapping process, in agreement with the Motor Theory of speech perception (Liberman & Mattingly, 1985). Furthermore, it has been proposed that the premotor cortex was a probable site for the mirror neuron system involved in the understanding of action (Rizzolatti & Craighero, 2003). In the context of the Motor Theory of speech perception, the mirror neuron system has been viewed as a possible link between sender and receiver (Rizzolatti & Arbib, 1998). Considering that, in oral communication, articulatory movements produce sounds that contribute to the verbal semantic content, the function of the mirror neuron system should be adjusted to auditory speech processing. Accordingly it has been suggested that speech production may be useful for perception in "challenging listening situations" (Lotto, Hickok, & Holt, 2009). Thus,

auditory-to-articulatory mapping would not be mandatory but might occur if need be to facilitate speech perception in adverse conditions, *i.e.* in a noisy environment. That being the case, one would have expected that the auditory contrast based on vocalic labiality, *i.e.* the most degraded in the fMRI environment of the present study, should have induced greater activation in the left premotor cortex than the other two phonological contrasts, more easily discriminated in this environment. As regards the greater effect of the fMRI noise on the vocalic contrast, it should be reminded that vowels and stop consonants further differ with respect to their degree of categorical perception (Eimas, 1963; Gerrits & Schouten, 2004). Indeed, the categorisation of phonemes is predicted by their ability to be classified with respect to each other. In contrast to stop consonants characterised by rapid acoustic changes and brief outbursts, vowels, which are more uniform over longer durations, are less precisely categorised. One may suggest that this intrinsic categorical difference could explain the specific involvement of the premotor cortex in the discrimination of contrasts relating to voicing or place of articulation; as opposed to contrast involving vocalic labiality. The significant increase revealed by the ROI analysis indicates a specific involvement of the left Premotor cortex in phonological processing. Yet, in the absence of significant difference between the static and dynamic AV modalities, this involvement does not appear to depend on speech movements, *per se*.

4.3. Temporal correlates of visemic processing

In order to discuss the ERP data of this study, one should bear in mind the particular context of the signal acquisition. The physical experimental constraints inherent in simultaneous fMRI–EEG data acquisition made it necessary to remove the pulse artefact, and the discrimination task was performed in a severely acoustically disturbed environment. Hence, ERP analyses focused on the identification of significant waveforms allowing comparisons across static and dynamic AV modalities. Additionally, unlike in other studies involving passive listening, detection or identification tasks, our syllabic discrimination task was not intended to investigate typical auditory potentials, but aimed at identifying the temporal and morphological changes related to the presentation of dynamic visual cues. Two significant waveforms peaking in the 125–175 ms and 225–275 ms temporal windows were identified.

First, a positive wave over Fz, Cz, Pz in the 125–175 ms window was recorded in response to the correct discrimination of syllables contrasting in respect of voicing, in the static AV modality, and in response to the correct discrimination of all three phonological contrasts, in the dynamic AV modality. Its amplitude significantly increased when the vocalic labiality-based contrast was discriminated in the dynamic *vs.* static AV modality. As mentioned before, this phonetic contrast was the most degraded by the fMRI environment, and its discrimination improved the most when the subjects were presented with full speech movements. This would suggest that the influence of visemic processing on syllabic discrimination in a noisy environment occurred as early as 150 ms following the speech stimulus onset.

Because of the very different experimental contexts, this finding cannot be directly compared to the significant decrease in amplitude of the N1 component of auditory ERPs in response to syllable identification in bimodal AV *vs.* unimodal stimulus presentation in an undisturbed environment, previously reported (Besle et al., 2004; Pilling, 2009; van Wassenhove et al., 2005). Our study did not evidence faster auditory speech processing. Indeed, in both conditions the syllable discrimination task involved AV stimuli, and improved behavioural performances, reported for the vocalic labiality features, only consisted in an increased accuracy rate without significantly reduced response time.

As to the significant negative wave recorded over the vertex sites in the 225–275 ms temporal window, elicited in response to the correct discrimination of the three phonological contrasts regardless of the modality of stimulus presentation, it was consistent with the N250 described by Vidal, Bonnet-Brilhault, Roux, and Bruneau (2005). During passive auditory perception of tones and syllables, these authors identified after the N1-P2 auditory potentials a negative wave sensitive to tone duration and speech stimuli, which they proposed as an index of speech stimulus processing. Moreover, in the present study, the amplitude of the so-called N250 increased significantly for the three contrasts in the dynamic modality (full articulatory configuration), suggesting a possible relationship with speech processing based on perceptuo-motor integration. This suggestion would be in agreement with the hypothesis that N250 might be a suitable candidate to study auditory perception in subjects with specific language-impairment (Vidal et al., 2005).

In line with an early study by Sumby and Pollack (1954), our results suggest that the more the auditory speech component is degraded, the more relevant congruent visual cues become for syllabic discrimination. This benefit from dynamic visual cues appears to occur as early as 150 ms post stimulus onset. Our fMRI data showed that the visual movement area (MT + V5) and the premotor cortex were involved in visemic processing. While the premotor cortex, a probable part of the mirror neuron system, was specifically activated by dynamic speech articulation, activation of the MT + V5 area did not depend on whether facial movements were associated or not with speech. The fact that activation in the premotor cortex does not rely on the visual distinctiveness of speech movement, and may be observed as a result of speech listening in absence of any visual cue (Wilson et al., 2004), suggests that its involvement could be related to motor speech representations rather than visemic processing *per se*. This view would be in accordance with theoretical frameworks advocating multimodal speech representations based on perceptuo-motor features, as postulated by the Perception-for-Action-Control Theory (Schwartz et al., in press).

Acknowledgements

We are grateful to N. Heider for administrative assistance and reviewing. We thank C. Marrer and G. Brock for technical assistance. We also thank the CIC of the Hôpitaux Universitaire de Strasbourg for examining the healthy volunteers. The work was supported of the French Ministry of Research (Grant ACI TTT, Mesures et Données, 2003–2006; ANR (DOCVACIM), 2008–2011) as well as Maison Interuniversitaire des Sciences de l'Homme d'Alsace (MISHA).

References

Abry, C., & Boë, L.-J. (1986). "Laws" for lips. *Speech Communication*, *5*(1), 97–104.
Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801.
Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). fMRI and its interpretations: An illustration on directional selectivity in area V5/MT. *Trends in Neurosciences*, *31*(9), 444–453.
Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*(5), 809–823.
Bernstein, L. E., Auer, E. T. J., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport*, *13*(3), 311–315.
Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234.
Blumstein, S. E., Cooper, W. E., Zurif, E. B., & Caramazza, A. (1977). The perception and production of voice onset time in aphasia. *Neuropsychologia*, *15*, 371–383.
Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox. *Presented at the 8th international conference on functional mapping of the human brain, June 2–6, Sendai, Japan.*
Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*, 299–320.

Buchsbaum, B., Pickell, B., Love, T., Hatrak, M., Bellugi, U., & Hickok, G. (2005). Neural substrates for verbal working memory in deaf signers: fMRI study and lesion case report. *Brain and Language*, *95*(2), 265–272.

Burton, M. W., Small, S. L., & Blumstein, S. E. (2000). The role of segmentation in phonological processing: An fMRI investigation. *Journal of Cognitive Neuroscience*, *12*(4), 679–690.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213–2218.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593–596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*(1), 57–70.

Campbell, R. (2008). The processing of audio–visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, *12*(2), 233–243.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2–3), 109–114.

Devlin, J. T., Raley, J., Tunbridge, E., Lanary, K., Floyer-Lea, A., Narain, C., et al. (2003). Functional asymmetry for auditory processing in human primary auditory cortex. *Journal of Neuroscience*, *23*(37), 11516–11522.

Eimas, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, 206–217.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*, 796–804.

Fowler, C. (1996). An event approach of the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Fridriksson, J., Moss, J., Davis, B., Baylis, G. C., Bonilha, L., & Rorden, C. (2008). Motor speech perception modulates the cortical language areas. *NeuroImage*, *41*(2), 605–613.

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T. E., & Penny, W. D. (Eds.). (2007). *Statistical parametric mapping: The analysis of functional brain*. London: Academic Press.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 535–609.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–376.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory–visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory–visual integration. *Journal of the Acoustical Society of America*, *103*(5), 2677–2690.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience*, *17*(6), 939–953.

Hasnain, M. K., Fox, P. T., & Woldorff, M. G. (1998). Intersubject variability of functional areas in the human visual cortex. *Human Brain Mapping*, *6*(4), 301–315.

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory–motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*, *15*(5), 673–682.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1–2), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Review Neuroscience*, *8*(5), 393–402.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*(1), 875–884.

Hutchison, E. R., Blumstein, S. E., & Myers, E. B. (2008). An event-related fMRI investigation of voice-onset time discrimination. *NeuroImage*, *40*(1), 342–352.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *Neuroreport*, *14*(8), 1129–1133.

Kiebel, S. J., & Friston, K. J. (2004a). Statistical parametric mapping for event-related potentials: I. Generic considerations. *NeuroImage*, *22*(2), 492–502.

Kiebel, S. J., & Friston, K. J. (2004b). Statistical parametric mapping for event-related potentials (II): A hierarchical temporal model. *NeuroImage*, *22*(2), 503–520.

Le Goff, B., & Benoît, C. (1996). A text-to-audiovisual-speech synthesizer for French. In *Proceedings of the international conference on spoken language processing, ICSLP 96* Philadelphia, (pp. 2163–2166).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lotto, A. J., Hickok, G., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110–114.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.

Marshall, J. C., & Fink, G. R. (2001). Spatial cognition: Where we were and where we are. *NeuroImage*, *14*(1), S2–S7.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Myatchin, I., Mennes, M., Wouters, H., Stiers, P., & Lagae, L. (2009). Working memory in children with epilepsy: An event-related potentials study. *Epilepsy Research*, *86*(2-3), 183–190.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*(1), 14–19.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*(2), 333–338.

Okada, K., & Hickok, G. (2006). Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. *Brain and Language*, *98*, 112–117.

Okada, K., & Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data. *Neuroscience Letters*, *452*(3), 219–223.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Otzenberger, H., Gounot, D., & Foucher, J. R. (2007). Optimisation of a post-processing method to remove the pulse artifact from EEG data recorded during fMRI: An application to P300 recordings during e-fMRI. *Neuroscience Research*, *57*(2), 230–239.

Pa, J., & Hickok, G. (2008). A parietal–temporal sensory-motor integration area for the human vocal tract: Evidence from an fMRI study of skilled musicians. *Neuropsychologia*, *46*(1), 362–368.

Peeters, M., Verhoeven, L., de Moor, J., & van Balkom, H. (2009). Importance of speech production for phonological awareness and word decoding: The case of children with cerebral palsy. *Research in Developmental Disabilities*, *30*(4), 712–726.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *Neuroreport*, *16*(2), 125–128.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech Language and Hearing Research*, *52*(4), 1073–1081.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1–2), 21–52.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience*, *21*(5), 188–194.

Rizzolatti, G., & Craighero, L. (2003). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, *45*(3), 587–597.

Saito, D. N., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M., & Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio–visual speech integration: A functional MRI study. *Cerebral Cortex*, *15*(11), 1750–1760.

Schwartz, J.-L., Abry, C., Boë, L.-J., & Cathiard, M.-A. (2002). Phonology in a theory of perception-for-action-control. In J. L. Durand, & B. Laks (Eds.), *Phonetics, phonology, and cognition*. Oxford University Press.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, in press.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neuroscience*, *26*(2), 100–107.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*(1–2), 13–45.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing? In H. Bouma, & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 165–182). NJ: Erlbaum.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory–visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*(3), 277–287.

Skipper, J. L., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor activation during speech perception. *NeuroImage*, *25*, 76–89.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387–2399.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–216.

Turkeltaub, P. E., & Branch Coslett, H. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, *43*(2), 271–282.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *PNAS*, *102*(4), 1181–1186.

Vidal, J., Bonnet-Brilhault, F., Roux, S., & Bruneau, N. (2005). Auditory evoked potentials to tones and syllables in adults: Evidence of specific influence on N250 wave. *Neuroscience Letters*, *378*(3), 145–149.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*, 989–994.
Wilms, M., Eickhoff, S., Specht, K., Amunts, K., Shah, N., Malikovic, A., et al. (2005). Human V5/MT+: Comparison of functional and cytoarchitectonic data. *Anatomy and Embryology*, *210*(5), 485–495.
Wilson, S. M., Molnar-Szakacs, I., & Iacoboni, M. (2008). Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cerebral Cortex*, *18*(1), 230–242.
Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702.